

There Are Approximately 1.5×10^{84} Possible Arrangements for the Biological Genetic Code

Douglas C. Youvan

doug@youvan.com

September 8, 2023

The genetic code is a marvel of nature, underpinning the very essence of life. Within its structure lie 64 codons that translate into 21 distinct entities, including amino acids and a stop signal. But have you ever wondered just how many unique arrangements this code can produce? This paper delves into that very question, merging the realms of biology and mathematics to uncover a staggering number: approximately 1.5×10^{84} . By navigating this vast landscape of possibilities, readers will gain a newfound appreciation for the intricacy of life's foundational code and the mathematical tools that help us decode its mysteries.

Keywords: genetic code, combinatorics, codons, amino acids, Inclusion-Exclusion Principle, mathematical biology, arrangements, sequence diversity.

Note: This calculation has been made many times before, but in this paper, GPT-4 was used to pick the applicable combinatorial math and wrote the Mathematica code, thus giving more confidence to this previously established number of $\sim 1.5 \times 10^{84}$.

In the vast realm of biology, the genetic code stands as a cornerstone of life. It is a set of rules that dictates how sequences of nucleotide triplets, called codons, specify which amino acid will be added next during protein synthesis. With 64 possible codons coding for 20 amino acids and one stop signal, understanding the genetic code's combinatorics offers a profound insight into the complexity of life at the molecular level.

This intricate dance of nucleotides and amino acids isn't just a concern for biologists; it presents a fascinating problem for mathematicians. The challenge? Determining the number of ways these codons can be arranged while ensuring that each of the 21 unique entities (20 amino acids + 1 stop signal) is represented at least once. It's a problem that seamlessly merges the worlds of biology and combinatorial mathematics.

Objective of the Paper:

This paper aims to bridge the understanding gap between two distinct audiences: those intrigued by the concept but lacking the technical background, and those deeply embedded in mathematical rigor. For the lay person, our goal is to provide an accessible insight into the genetic code's complexity and why such problems matter. For the mathematician, we present a thorough exploration of the combinatorics behind the genetic code, employing the inclusion-exclusion principle to elucidate the solution. By offering dual perspectives, we hope to foster a deeper appreciation of the beauty and intricacy of life and the mathematical tools we use to understand it.

For the Lay Person

Understanding the Genetic Code: A Simple Analogy

Imagine walking into a vast library. This isn't an ordinary library but one that contains the stories of every living organism on Earth, from the smallest bacteria to the blue whale, even humans. Each book in this library corresponds to an organism, and the text within tells its life story.

Now, think of the words in these books as proteins. These proteins define the characteristics of that organism—whether a tree has leaves or needles, whether a bird can fly or swim. But how are these words (proteins) formed?

Here comes the magic: instead of 26 letters as in the English alphabet, life uses just 21 'letters' (20 amino acids + 1 stop signal) to form its words. And each 'word' is made up of a combination of three 'letters', giving us a vocabulary of 64 possible words (codons).

However, the real challenge lies in creating sentences (sequences) using all 64 words at least once, without any restrictions on repetition. This puzzle, which may sound like a game, holds the key to understanding the vast diversity and complexity of life.

Breaking Down the Numbers: Why Does it Matter?

You might wonder, with such a vast universe, why does the arrangement of these 'words' in a 'sentence' matter? Well, the sequence in which these codons are arranged directly influences the sequence of amino acids in a protein, which in turn defines its function and structure. A slight change, and you might have a protein that gives a firefly its glow or one that aids in human digestion.

Considering the combinatorial possibilities ($\sim 1.51 \times 10^{84}$ ways) is like pondering the number of unique stories you could write with a 64-word vocabulary. It's a testament to the immense diversity and adaptability of life. Understanding these numbers isn't just a mathematical feat; it's a journey into the depth and breadth of possibilities that nature offers.

Moreover, by grasping the scale of these numbers, we appreciate the meticulous precision of biological processes. Errors are rare, given the vast number of potential combinations, emphasizing the marvel that is life's coding mechanism.

For the Mathematician

Setting Up the Problem: Counting with Constraints:

At the heart of this problem is a fundamental question of combinatorial mathematics: given 64 objects (codons) of 21 different types (20 amino acids + 1 stop signal), with at least one object of each type and indistinguishability within its type, how can we count the ways to arrange these objects in a sequence?

The problem might remind many of classic combinatorics problems, such as distributing balls into urns, with the added constraint that each urn must have at least one ball. In our case, each 'urn' (amino acid or stop signal) can be represented multiple times in the sequence, but it must appear at least once.

Applying the Inclusion-Exclusion Principle:

The Inclusion-Exclusion Principle provides a systematic method to count sets by considering all unrestricted counts first and then adjusting for overcounts by subtracting, adding, subtracting, and so on.

To apply it to our problem:

1. We first count sequences without restrictions: 21^{64} .
2. Then we account for sequences that exclude at least 1 of the 21 types: subtracting $\binom{21}{1} \times 20^{64}$.
3. Add back the sequences that exclude 2 of the 21 types: adding $\binom{21}{2} \times 19^{64}$.
4. Subtract sequences excluding 3 types, add those excluding 4, and so on, until we reach sequences missing 20 of the 21 types.

Deriving the Solution: Step-by-Step Calculation:

Given the previous application of the Inclusion-Exclusion Principle, the number of sequences that include at least one of each type can be derived as:

$$21^{64} - \binom{21}{1} \times 20^{64} + \binom{21}{2} \times 19^{64} - \dots + \binom{21}{20} \times 1^{64}$$

Computing this expression provides the astonishing result of approximately 1.51×10^{84} ways. This number isn't just massive; it's on par with quantities often reserved for cosmology or quantum mechanics. It represents the immense variety of sequences that nature can draw from when crafting proteins, highlighting the power of combinatorial mathematics in elucidating biological mysteries.

Implications and Interpretations:

At its core, the number 1.51×10^{84} stands testament to the vast potential intricacies hidden within the genetic code. A number so vast it dwarfs even the estimated count of atoms in the observable universe, which stands around 10^{80} . But beyond its

sheer magnitude, this number carries deeper implications for both biology and mathematics.

For Biology: The wide array of potential sequences reflects the dynamism and adaptability inherent in life. With such an expansive canvas, it's no wonder we find the immense biodiversity on our planet. These combinations give rise to various proteins, each with unique functionalities, allowing organisms to adapt, evolve, and thrive in diverse environments. It underscores the idea that even with a set number of building blocks, life's potential configurations are near boundless, providing a reservoir for evolution to draw from and adapt.

For Mathematics: This exercise serves as a prime example of the power and utility of combinatorial techniques. The Inclusion-Exclusion Principle, a tool rooted in abstract mathematical theory, finds a very tangible application, revealing insights into the tangible world. It's a reminder that mathematics often provides the language and tools to dissect and understand complex phenomena, bridging seemingly disparate domains.

In conclusion, the genetic code's combinatorics is more than a numerical exercise. It's a celebration of life's complexity and the beauty of mathematical structures that allow us to decode it. Both laypersons and mathematicians can marvel at this interplay, finding awe in the vastness of possibilities and the precision of nature's coding mechanism.